Genomics and proteomics: a signal processor's tour

P. P. Vaidyanathan, Fellow, IEEE

Contact Author: P. P. Vaidyanathan,[1] Dept. Electrical Engr., 136-93, California Institute of Technology, Pasadena, CA 91125. Ph:(626) 395 4681 Email: ppvnath@systems.caltech.edu

Invited article, IEEE Circuits and Systems Magazine, March 2005. Based on a plenary lecture given at the IEEE International Symposium on Circuits and Systems, May 2004.

Abstract. The theory and methods of signal processing are becoming increasingly important in molecular biology. Digital filtering techniques, transform domain methods, and Markov models have played important roles in gene identification, biological sequence analysis, and alignment. This paper contains a brief review of molecular biology, followed by a review of the applications of signal processing theory. This includes the problem of gene finding using digital filtering, and the use of transform domain methods in the study of protein binding spots. The relatively new topic of noncoding genes, and the associated problem of identifying ncRNA buried in DNA sequences are also described. This includes a discussion of hidden Markov models and context free grammars. Several new directions in genomic signal processing are briefly outlined in the end.

Keywords. *Genomic-signal-processing, bioinformatics, genes, protein-coding, DNA, and ncRNA.*

1. INTRODUCTION

Subsequent to the sensational announcement of the double helix structure for the DNA molecule more than fifty years ago [1], there has been phenomenal progress in genomics in the last five decades. With the enormous amount of genomic and proteomic data available to us in the public domain, it is becoming increasingly important to be able to process this information in ways that are useful to humankind. Traditional as well as modern signal processing methods have played an important role in these fields. Genomic signal processing is primarily the processing of DNA sequences, RNA sequences, and proteins. A DNA sequence is made from an alphabet of four elements, namely A, T, C, and G. For example

$$...ATCCCAAGTATAAGAAGTA...$$

The letters A, T, C, G represent molecules called **nuclotides** or **bases** (to be described soon). Since DNA contains the genetic information of living organisms, we see that life is governed by quarternary codes.

[1]Work supported in part by the ONR grant N00014-99-1-1002.

Another example of discrete-alphabet sequences in life forms is the protein. A large number of functions in living organisms are governed by proteins. A protein can be regarded as a sequence of **amino acids**. There are twenty distinct amino acids, and so a protein can be regarded as a sequence defined on an alphabet of size twenty. The twenty letters used to denote the amino acids are the letters from the English alphabet except B, J, O, U, X, and Z. For example a part of the protein sequence could be

$$...PPVACATDEEDAFGGAYPQ...$$

Notice that some letters representing amino acids are identical to some letters representing bases. For example the A in the DNA is a base called adenine, and the A in the protein is an amino acid called alanine.

If we assign numerical values to the four letters in the DNA sequence, we can perform a number of signal processing operations such as Fourier transformation [26,3], digital filtering [27], time-frequency plots such as wavelet transformations [17], and Markov modelling [4]. Some of those are quite interesting and in fact have important practical applications. Similarly, once we assign numerical values to the twenty amino acids in protein sequences we can do useful signal processing.

Scope and outline

This magazine article is meant only to be an introduction. The aim here is to present a big picture with appropriate background information. The field is quite mature, and the reader with serious interest should pursue some of the references cited at the end of this article. For convenience the references are categorized by topic.

Sections 2—5 contain brief but important background material on DNA and proteins sequences. In Sec. 6 we explain how Fourier techniques have played a role in gene identification and protein analysis. Section 7 explains the role of hidden Markov models in molecular biology. We then discuss in Sec. 8 noncoding genes which have been increasingly recognized for their important role in nearly all life forms. A brief overview of issues involved in computational identification of noncoding genes is also presented. We conclude the paper with further remarks on topics of recent interest. Overviews of some of the important aspects of genomic signal processing can be found in the introductory magazine-article by Anastassiou [3] and in a recent journal article [8].

2. SOME FUNDAMENTALS

Figure 1(a) shows a schematic for the DNA (deoxyribo nucleic acid) molecule. This is in the form of a double helix. The discovery of this double helix is one of the landmarks of molecular biology (for detailed story, see Box B1). Between the two strands of the backbone which is outside, there are pairs of bases like the rungs of a ladder. The backbone is a very regular structure made from sugar-phosphate. There are four types of

bases (or nucleotides), denoted with the letters A, C, G, and T (respectively, adenine, cytosine, guanine, and thymine). For completeness, the internal atomic details of the molecules A, T, C, and G are shown in Fig. 2. These molecules are made from carbon, nitrogen, hydrogen and oxygen atoms. There are about three billion of these bases in the DNA of a single human cell (Fig. 3).

In Fig. 1(b) the double helix is shown straightened out for simplicity. The genome sequence corresponding to the top strand of the DNA molecule in this example is $AGACTGAA$. Note that the ordering is from the so-called $5'$ to the $3'$ end (left to right). DNA sequences are typically listed from the $5'$ to the $3'$ end because they are scanned in that direction when bases are used by the cell machinery to signal the production of amino acids. The reason for directed flow arises from the way the sugar and phosphate are glued together (Fig. 1(c)). In the double stranded DNA, the base A always pairs with T, and C pairs with G. Thus the bottom strand $TCTGACTT$ is the complement of the top strand. This **base-pairing** occurs through a weak bond called the **hydrogen bond** [2] but because there are several million base pairs, the two strands are held together strongly. Typically in any given region of the DNA molecule, at most one of the two strands is active in gene expression (Sec. 3).

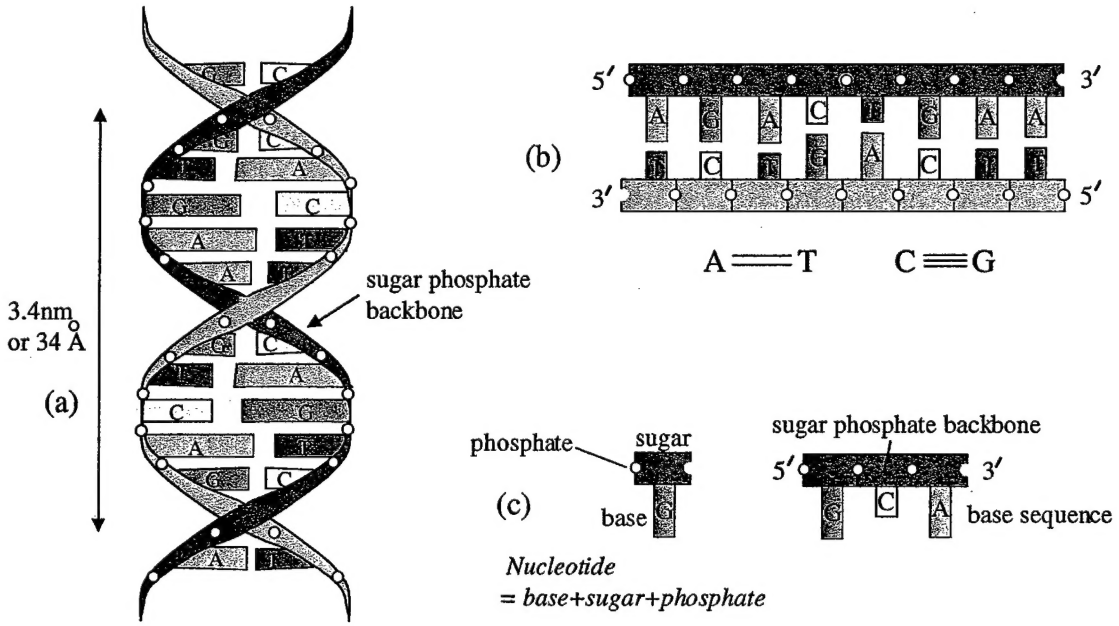


Figure 1. (a) The DNA double helix, (b) linearized schematic, and (c) details of the the sugar-phosphate backbone. In part (b) bottom strand is complementary to the top strand in the sense that A and T are paired and so are C and G. This is because of a weak bonding called hydrogen bonding between these pairs of molecules.

Single-celled organisms like bacteria do not have a nucleus and the DNA just resides in the cell. Such cells are called **procaryotes**; higher organisms (worms, insects, plants, mammals, ...) have cells with nucleus and

are called **eucaryotes**. These have the DNA residing in the nucleus. An exception is the red blood cell which has no nucleus. Cells also have a small quantity of DNA in the mitochondria; we shall not discuss this here.

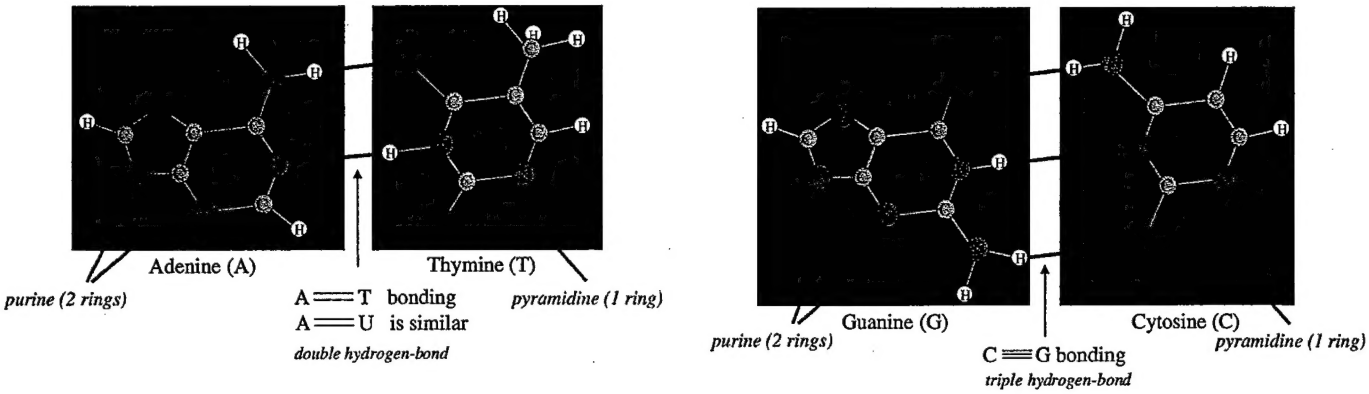


Figure 2. Internal atomic details of the bases adenine, thymine, guanine, and cytosine. These molecules are made from carbon, hydrogen, oxygen and nitrogen (hence called *nitrogenous* bases). Note that A and G have two rings and are called *purines*. The molecules C and T have one ring and are called *pyramidines*.

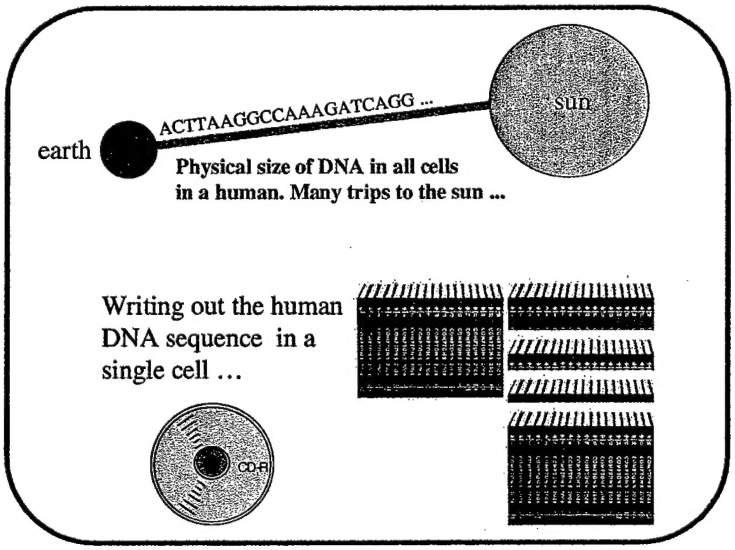


Figure 3. *A feeling for sizes...* The DNA in the nucleus of a single human cell is about 3 *billion* bases long (and is organized into 46 chromosomes). For typical bacteria the DNA is about 4 *million* bases long. If the DNA in a human cell is stretched out like a piece of string, it stretches out to 2 yards! If we put together all DNA in all the (5 trillion) cells in an average human, the length is sufficient to cover the distance from earth to the Sun (93 million miles), about 50 times. If we were to write down each base using normal letter size, the DNA in a single human cell would fill about 2000 novels. If the three billion bases in a human genome are stored digitially using two bits to code each base location (of four possible bases), the total is 6 billion bits or equivalently 750 Mega bytes (roughly the capacity of a standard CD). A typical cell nucleus which is one hundredth of a millimeter across can store as much information as does a CD!

The **RNA** (ribo nucleic acid) molecule is closely related to the DNA. It is also made of four bases but instead of thymine, a molecule called **uracil** is used (denoted as U). The sugar in the sugar-phosphate backbone

is also slightly different but we do not require the details here. The important fact is that U pairs with A by hydrogen bonding just like T pairs with A. RNA molecules are short (and typically short-lived) single-stranded molecules which are used by the cell as temporary copies of portions of DNA (Sec. 3).

3. GENES AND DNA

A DNA sequence can be separated into two types of regions: **genes** and **intergenic** spaces. Genes contain the information for generation of proteins. Each gene is responsible for the production of a different protein as shown schematically in Fig. 4. Even though all the cells in an organism have identical genes, only a selected subset is active in any particular family of cells. For example the set of genes that are active in blood cells are different from those that are active in nerve cells, which explains why these cells look so different! See Fig. 5.

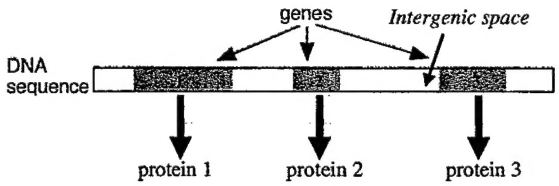


Figure 4. Genes are parts of the DNA sequence, and are responsible for the production of proteins. According to classical view (central dogma of biology) each gene produces a specific protein. See text.

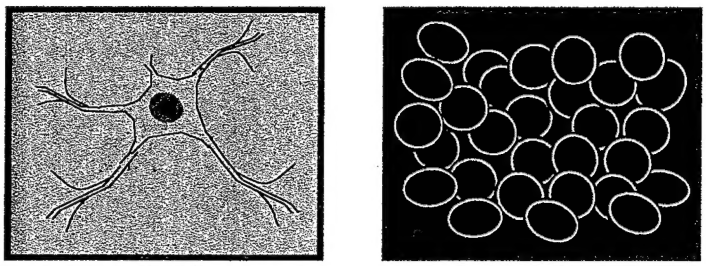


Figure 5. Brain cells and red blood cells. Cells look very different from each other because of the different sets of genes expressed in them. See www-biology.ucsd.edu/news/article_112901.html and www.cellsalive.com/ gallery.htm for real micrographs.

Figure 6 shows some of the steps involved in the production of a protein from a gene. Notice that a gene has two types of subregions called the **exons** and **introns** (procaryotes like bacteria do not have introns).[2] The gene is first copied into a single stranded chain called the messenger RNA or **mRNA** molecule. The introns are then removed from the mRNA by a process called splicing. The spliced mRNA is then used by a large molecule called the **ribosome** to produced the appropriate protein. The translation from mRNA to protein is aided by adaptor molecules called the transfer RNA or **tRNA** molecules. In some sense the tRNA molecules

[2]The existence of introns came to the attention of the scientific community only in 1977 [2].

store the genetic code as we shall see in Sec. 4. Ribosomes are often referred to as the protein factories of the cell. There are many ribosomes in a cell working in parallel like molecular machines.

Many details are omitted in Fig. 6 for brevity. For example the mRNA is in reality the *complement* of the gene, that is, Cs are replaced with Gs, and As with Ts (rather Us). Thus, if the gene is $ATTAGC$ then the mRNA is $UAAUCG$. There is a second level of complementing which cancels this when the mRNA attaches to tRNA molecules at the so-called anticodon sites.

The observation that each gene is responsible for the creation of a protein (through mRNA) is often expressed as

$$\text{gene in } DNA \rightarrow RNA \rightarrow protein$$

and is referred to as the **central dogma** of molecular biology. We will see in Sec. 8 that the dogma has been challenged in recent years.

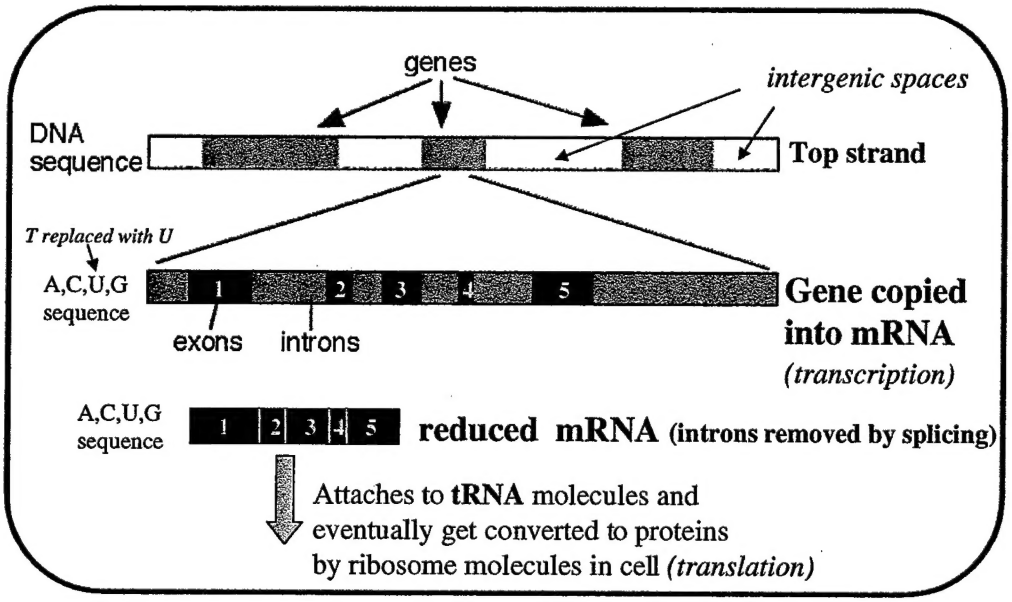


Figure 6. When a gene is ready to be expressed, it is duplicated in the form of a single-strand molecule called the **mRNA** (messenger RNA) which then leaves the nucleus. The introns are spliced out and a shorter mRNA molecule is produced. Thus, unlike the parent gene, the mRNA is a concatenation of the exons only. It is used by ribosomes outside the nucleus of the cell to manufacture the appropriate protein coded by the original gene. Thus protein production involves the *transcription* of genes into mRNA and the subsequent *translation* of the 4-letter language to a 20-letter language.

4. THE GENETIC CODE

How does the cell know what protein to make from a particular gene? This information is contained in a code which is common to all life. Recall that the gene gets duplicated into the mRNA molecule which is then spliced so that it contains only the exons of the gene. Imagine that this spliced mRNA is divided into groups

of three adjacent bases. Each triplet is called a **codon**. Evidently there are 64 possible codons. Thus the mRNA is nothing but a sequence of codons. Each codon instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein. This mapping is called the genetic code and is shown[3] in Fig. 7. Since there are 64 possible codons but only 20 amino acids, the mapping from codons to amino acids is many-to-one (Fig. 8). The story of how the genetic code was cracked is summarized in Box B2.

AAA: K (Lys)	*GAA* : E (Glu)	*TAA* : Stop	*CAA* : Q (Gln)
AAG: K (Lys)	*GAG* : E (Glu)	*TAG* : Stop	*CAG* : Q (Gln)
AAT: N (Asn)	*GAT* : D (Asp)	*TAT* : Y (Tyr)	*CAT* : H (His)
AAC: N (Asn)	*GAC* : D (Asp)	*TAC* : Y (Tyr)	*CAC* : H (His)
AGA: R (Arg)	*GGA* : G (Gly)	*TGA* : Stop	*CGA* : R (Arg)
AGG: R (Arg)	*GGG* : G (Gly)	*TGG* : W (Trp)	*CGG* : R (Arg)
AGT: S (Ser)	*GGT* : G (Gly)	*TGT* : C (Cys)	*CGT* : R (Arg)
AGC: S (Ser)	*GGC* : G (Gly)	*TGC* : C (Cys)	*CGC* : R (Arg)
ATA: I (Ile)	*GTA* : V (Val)	*TTA* : L (Leu)	*CTA* : L (Leu)
ATG: M (Met)	*GTG* : V (Val)	*TTG* : L (Leu)	*CTG* : L (Leu)
ATG = Start			
ATT: I (Ile)	*GTT* : V (Val)	*TTT* : F (Phe)	*CTT* : L (Leu)
ATC: I (Ile)	*GTC* : V (Val)	*TTC* : F (Phe)	*CTC* : L (Leu)
ACA: T (Thr)	*GCA* : A (Ala)	*TCA* : S (Ser)	*CCA* : P (Pro)
ACG: T (Thr)	*GCG* : A (Ala)	*TCG* : S (Ser)	*CCG* : P (Pro)
ACT: T (Thr)	*GCT* : A (Ala)	*TCT* : S (Ser)	*CCT* : P (Pro)
ACC: T (Thr)	*GCC* : A (Ala)	*TCC* : S (Ser)	*CCC* : P (Pro)

Figure 7. The genetic code. Triples of bases such as AAA denote codons. The single letters such as K denote amino acids. Their three letter names (e.g., Lys) are also shown. Full names of amino acids can be found in Fig. 8.

When a gene is expressed, each codon in the mRNA produces an amino acid according to the genetic code, and the amino acids are bonded together into a chain. Figure 9 shows an example of how mRNA is coverted to protein using the genetic code. When all the codons in the mRNA are exhausted we get a long chain of amino acids (typically a few hundred long). This is the protein corresponding to the original gene. Notice that there is a **start codon** ATG which signifies the beginning of the protein-coding part of the gene. If a start codon occurs inside a gene again, it produces the amino acid methionine. A **stop codon** signifies that the protein coding part of the gene has come to an end. There are three stop codons. The chemical bond between amino acids is a *covalent peptide bond.* Figure 10 shows examples of two amino acids and the resulting bond.

[3]We have used T instead of U because the original gene has T. In fact We will use U and T rather interchangably; the context will make the distinction clear.

The translation of the codons into amino acids is made physically possible by adaptor molecules called transfer RNA or **tRNA** molecules. There are more than 20 kinds of tRNA molecules in the cell (at least one for each amino acid). One end of the molecule matches a specific codon and the other end attaches to the corresponding amino acid. See Fig. 11. The molecule ribosome (Sec. 3) works in conjunction with tRNA molecules and mRNA to produce the protein. So it is clear that the genetic code is essentially stored in the tRNA molecules.

1	*A*	Ala	Alanine	GCA,GCC,GCG,GCT
2	*C*	Cys	Cysteine (has *S*)	TGC, TGT
3	*D*	Asp	Aspartic acid	GAC,GAT
4	*E*	Glu	Glutamic acid	GAA,GAG
5	*F*	Phe	Phenylalanine[1]	TTC,TTT
6	*G*	Gly	Glycine	GGA,GGC,GGG,GGT
7	*H*	His	Histidine[2]	CAC,CAT
8	*I*	Ile	Isoleucine[3]	ATA,ATC,ATT
9	*K*	Lys	Lysine[4]	AAA,AAG
10	*L*	Leu	Leucine[5]	TTA,TTG,CTA,CTC,CTG,CTT
11	*M*	Met	Methionine[6] (has *S*)	ATG
12	*N*	Asn	Asparagine	AAC,AAT
13	*P*	Pro	Proline	CCA, CCC, CCG,CCT
14	*Q*	Gln	Glutamine	CAA,CAG
15	*R*	Arg	Arginine[7]	AGA,AGG,CGA,CGC,CGG,CGT
16	*S*	Ser	Serine	AGC,AGT,TCA,TCC,TCG,TCT
17	*T*	Thr	Threonine[8]	ACA,ACC,ACG,ACT
18	*V*	Val	Valine[9]	GTA,GTC,GTG,GTT
19	*W*	Trp	Tryptophan[10]	TGG
20	*Y*	Tyr	Tyrosine[11]	TAC,TAT

Figure 8. A list of the twenty amino acids, and codons which generate them (from Fig. 7). For example the amino acid alanine (A) can be generated by any one of four possible codons GCA, GCC, GCG, or GCT. The superscipts 1 to 11 indicate the eleven *essential amino acids* (some references say there are fewer than eleven). These by definition are the amino acids animals cannot manufacture — they need to eat them. Milk provides all essential amino acids, and so does a combination of grains and beans.

It is a wonder of Nature that all life forms (from bacteria to mammals) use the same genetic code. This is no doubt due to the common origin of all life. Can one change Nature's genetic code? Apparently this is not impossible. Recall from Fig. 7 that the stop codon TAG produces no amino acid. In 2001 Wang and Schultz added enough biological machinery in E. coli bacteria to enable it to synthesize a *new* amino acid from TAG. In 2003 they showed how this amino acid can be inserted in a E. coli protein made with natural amino acids. The same idea was successful in yeast. It has been suggested by some authors that such new proteins could be the key to destroying cancerous cells quickly. A Scientific American article which appeared in May 2004 describes some of these areas of research [44].

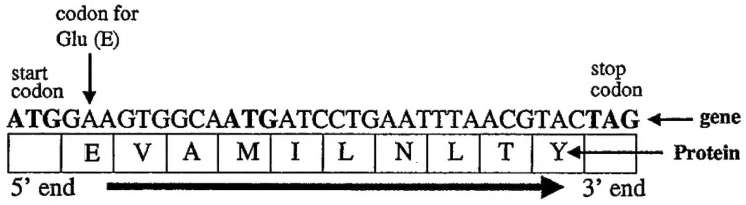


Figure 9. A toy example showing how a sequence of codons gets translated to a protein, ten amino-acids long. In most cases genes are much longer (thousands of bases); proteins have several hundred amino acids. Notice that if a base is deleted by accident somewhere in the middle, then *all* the codons following that point are changed, possibly changing all the amino acids following. If an entire codon is deleted, it is like deleting an amino acid; nothing else changes.

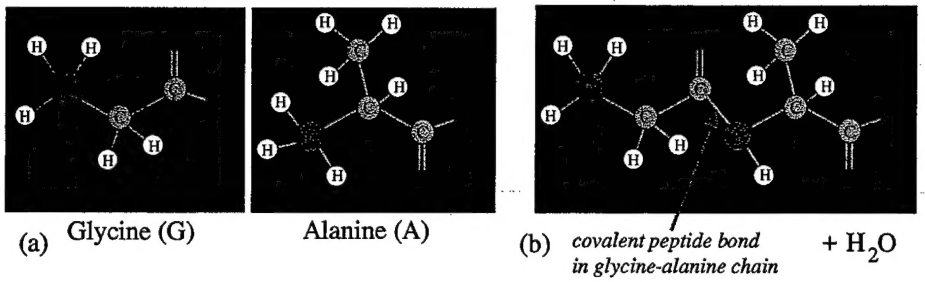


Figure 10. (a) Examples of two amino acids, and (b) bonding of these two amino acids, with consequent release of a water molecule. Like bases, amino acids are also made from carbon, hydrogen, oxygen, and nitrogen. Some of them also have sulfur (as indicated in Fig. 8).

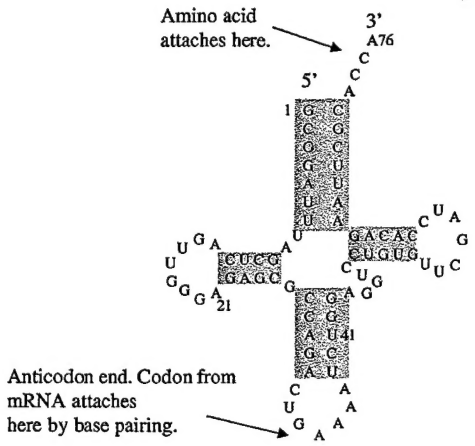


Figure 11. Example of a transfer RNA molecule in yeast. The bases are numbered from 1 to 76. Only a particular codon can match perfectly with the anticodon, and can therefore be associated with the specific amino acid that is able to attach to the tRNA at the top end. In this manner, the tRNA molecules store the genetic code in the cell.

5. PROTEINS

Because of the innumerable combinations from the alphabet of 20 amino acids, the number of different proteins in living organisms is enormous. Proteins drive most of the biological processes in living organisms. **Enzymes**, for example, are proteins with a special role, namely the speeding up of biochemical reactions in living organisms. Fibers in tendons and ligaments, components of hemoglobin (oxygen carrier in red blood cells), myosin in muscle cells (motor protein), ferritin in the liver, rhodopsin in retina (light detector) and harmones such as insulin, gastrin, and glucagon, are all proteins. When a protein is left in a watery medium it automatically *folds into a specific three dimensional structure*, which depends almost entirely on the amino acid sequence defining the protein (the pH or acidity of the watery medium is also important). The 3D shape of a protein allows it to interact only with very specific molecules in the cell, and this is important for the proper functioning of proteins.[4] In fact **protein folding** is a major area of research by itself. For example, given the amino acid sequence alone, can we predict the 3D folded shape using physics and mathematics alone? Figure 12 shows a computer drawing of the protein hemoglobin which is made of four smaller proteins [2]. Like DNA, proteins are macro molecules. The average protein is about 40,000 times heavier than a hydrogen atom. We will say more about the signal processing aspects in Sec. 6.3.

The discovery of the double helix also solved another mystery of molecular biology: it suggested how the huge **DNA is replicated** accurately in cell devision. Namely, the double strand separates or **unzips** into two single strands each of which serves as a mold to form a new complementary strand. (The unzipping process is also present locally when a gene is copied into an mRNA (Fig. 13)). Each single strand quickly manufactures the complementary strand from bases floating around in the cell. This was later verified by Matt Meselson and Frank Stahl, sometime after 1954, in an experiment considered to be one of the most beautiful experiments in biology. The accuracy of duplication is phenomenal because of the self error correcting mechanism (called mismatch-pair system) implicit in the cell [2]. The **probability of error** is about 10^{-9}. Compare this to a average typist (1 error per typped page) or the postal system (10 late deliveries out of every ...). Such accuracy is necessary in gene reproduction because even small changes in the DNA (mutations, insertions, deletions) can change the proteins made by the genes dramatically. For example, sickle cell anemia is created because of a single error in a gene (see Fig. 14). On the other hand there are examples where even multiple errors do not change the protein (because the codon to amino acid mapping has redundancy, Fig. 8. So the cell has built-in tolerance for errors; the example of sickle cell anemia is rather unusual.

[4]For example the enzyme thrombin reacts only with the protein fibrinogen (which is part of the blood clotting process). There are exceptions too: the digestive enzymes pepsin and chymotrypsin act on almost any protein they encounter. The Encyclopedia Britannica contains a wealth of information on this topic.

Figure 12. Pasta dish? No, it is an example of a protein (Hemoglobin, human). Figure taken from the website www.biochem.szote.u-szeged.hu/astrojan/protein2.htm, generated by the program MOLMOL (Koradi et al., 1996). See reference [15].

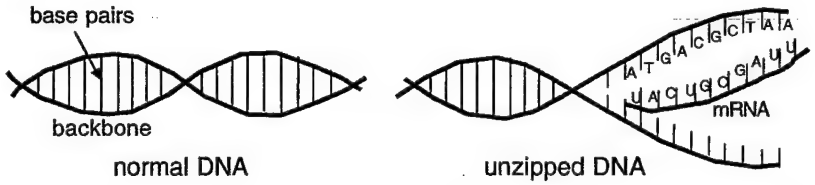


Figure 13. Unzipping of a DNA sequence to produce an mRNA copy of a selected region. This occurs during gene expression. Note that the mRNA strand is complementary to the DNA, that is A is replaced with a T (rather U which is similar) and vice versa; similarly C is replaced with G and vice versa. A similar unzipping separates the two DNA strands completely during cell division.

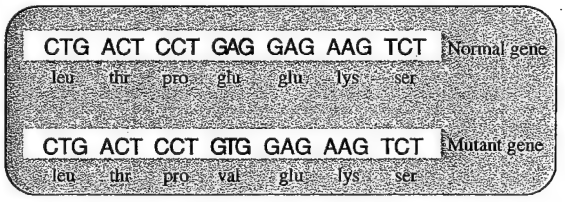


Figure 14. Cause of sickle-cell anemia. A gene called HBB in human chromosome 11 creates the protein beta globin in the hemoglobin of red blood cells. This gene is 1600 bases long. A single mutation (or base-change) in this gene gives rise to sickle-cell anemia. The figure shows portions of the normal gene and mutated gene. The codon GAG is changed to GTG, which means that the amino acid changes from glutamic acid to valine. This single change in the amino acid chain makes a crucial corner of the 3D protein molecule *hydrophobic* (water hating), and causes hemoglobin molecules to stick together and create rigid fibres.

6. FILTERING AND TRANSFORM-DOMAIN METHODS IN GENOMICS AND PROTEOMICS

The application of Fourier transform techniques has been found to be very useful both for DNAs and protein sequences. First it is convenient to introduce *indicator sequences* for bases in DNA. For example the

indicator for base A is a binary sequence of the form $x_A(n) = 000110111000101010\ldots$, where 1 indicates the presence of an A and 0 indicates its absence. The indicator sequences for the other bases are defined similarly. Denote the discrete Fourier transform [64] or DFT of a length-N block of $x_A(n)$ as $X_A[k]$, that is, $X_A[k] = \sum_{n=0}^{N-1} x_A(n)e^{-j2\pi kn/N}$, $0 \le k \le N-1$. The DFTs $X_T[k], X_C[k]$, and $X_G[k]$ are defined similarly.

6.1. Identifying Protein Coding Genes

It has been noticed that protein-coding regions (exons) in genes have a *period-3 component* because of coding biases in the translation of codons into amino acids. This observation can be traced back to the 1980 work of Trifonov and Sussman [35]. The period-3 property is not present outside exons, and can be exploited to locate exons. [3, 26]. Thus if we take N to be a multiple of 3 and plot

$$S[k] \triangleq |X_A[k]|^2 + |X_T[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 \tag{1}$$

then we should see a peak at the sample value $k = N/3$ (corresponding to $2\pi/3$). Given a long sequence of bases we can calculate $S[N/3]$ for short windows of the data, and then slide the window. Thus, we get a picture of how $S[N/3]$ evolves along the length of the DNA sequence. It is necessary that the window length N be sufficiently large (typical window sizes are a few hundreds, eg., 351, to a few thousands) so that the periodicity effect dominates the background $1/f$ spectrum (Sec. 6.2). However a long window implies longer computation time, and also compromises the base-domain resolution in predicting the exon location.

The sliding window method can be regarded as digital filtering followed by downsampling (at a rate depending on the separation between adjacent positions of the window [67]). The filter has a simple impulse response

$$w(n) = \begin{cases} e^{j\omega_0 n} & 0 \le n \le N-1 \\ 0 & \text{otherwise.} \end{cases}$$

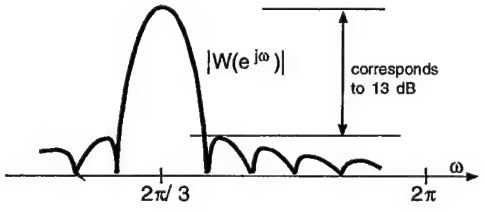


Figure 15. Computation of DFT with a sliding window is equivalent to lowpass digital filtering. The frequency response magnitude is as shown, and offers about 13 dB stopband attenuation.

This is a bandpass filter with passband centered at $\omega_0 = 2\pi/3$ and minimum stopband attenuation of about 13 dB (Fig. 15). If we pay careful attention to the design of the digital filter, we can isolate the period-3

behavior from background information such as $1/f$ noise more effectively. We can also use efficient methods to design and implement the filter, thereby reducing computational complexity.

Based on these observations, a number of methods have been proposed for designing digital filters suited to gene prediction application [27], [28]. We show in Fig. 16 the exon prediction results for gene F56F11.4 in the C. elegans chromosome III. This gene has five exons. The first plot uses the DFT based spectrum using a sliding window. The five peaks corresponding to the exons can be seen clearly. The second plot uses a multistage filter $H(z)$ similar to the IFIR filter advanced by Neuvo et al. [63]. Notice that the background noise (due to $1/f$ behavior, Sec. 6.2) has been removed almost completely and the five exons can be seen clearly. Further design details of this can be found [28] and in a recent tutorial article [8].

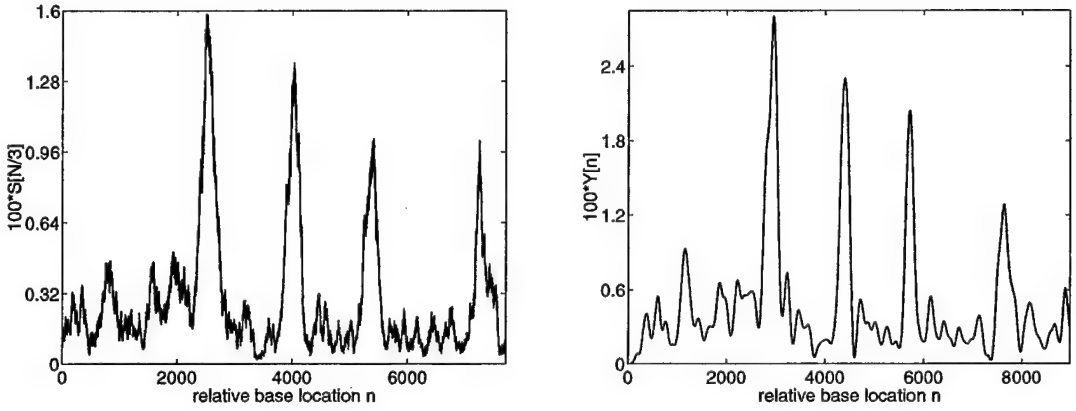


Figure 16. Left plot: the DFT based spectrum $S[N/3]$ for gene F56F11.4 in the C. elegans chromosome III. Right plot: the multistage narrowband bandpass filter output [28] for the same gene. The mulitstage filter does a very good job of eliminating the $1/f$ component in the DNA spectrum, and the exon regions are revealed more clearly.

Some authors have claimed that the period-3 property is due to nonuniform codon usage, also known as codon bias: even though there are several codons which code a given amino acid (Fig. 8), they are not used with uniform probability in organisms. For example base G dominates at certain codon positions in the coding regions [31]. We have, in fact, observed experimentally that the use of the plot $|X_G(k)|^2$, which depends on *base G alone*, is often quite sufficient for revealing the period-3 property, and therefore for the prediction of protein coding regions.

Does the method always work? Tiwari, et al. [26] have observed that some genes *do not* exhibit period-3 behavior at all in S. cerevisiae. Furthermore for procaryotes (cells without a nucleus), and some viral and mitochondrial base sequences, such periodicity has even been observed in noncoding regions [33]. For this and many other reasons [23], gene identification is a very complex problem, and the identification of period-3 regions is only a step towards gene and exon identification. Hidden Markov models (Sec. 7) have been used

quite successfully for this application [24].

6.2. Long range correlations or 1/f behavior

The period-3 behavior described above indicates strong short-term correlation in the coding regions. But there is also long-range correlation exhibited by DNA sequences both in the gene regions and intergenetic regions. One of the earliest papers to point this out appeared in *Nature* in 1992 [34]. The study was made based on a concept called the *DNA walk*. Latter studies by other authors examined correlations over much longer regions which contained many genes. Long range correlations have been found both in coding and noncoding regions [39]. According to Fourier transform theory, long range correlation implies that the Fourier transform has $1/f$-behavior in low frequency regions [65].

Another early work on the topic was the 1992 paper by Richard Voss [37] who was perhaps also the first person to define indicator sequences for bases, and calculate the deterministic autocorrelation. For example, letting $x_A(n)$ be the indicator for base A, the autocorrelation is $r_A(k) = \sum_n x_A(n)x_A(n-k)$, and the Fourier transform $S_A(e^{j\omega})$ of this is the power-spectrum for base A. Notice that $S_A(e^{j\omega}) = |X_A(e^{j\omega})|^2$. Voss analyzed the human Cytomegalovirus strain AD169. The genome length was $N = 229,354$. The lowest meaningful frequency[5] can be regarded as $1/N$ which is slightly smaller than $0.5 * 10^{-5}$. Voss demonstrated that the power spectrum has power-law or $1/f^{\beta}$ behavior for each of the four indicator sequences (for appropriate β close to unity). Later studies have indicated that such long range correlation is valid even further, extending to several millions of bases [36] (i.e., the $1/f$ behavior extends to even smaller frequencies). Figure 17 shows the power spectrum $S_A(e^{j\omega})$ for base A for the first one-million bases of an entire bacterial genome of length about 1.55 million. The organism is called *Aquifex aeolicus*, and its genome can be found in public websites such as the gene bank [29]. There were 0.5 million samples of $S_A(e^{j\omega})$ in $0 \leq \omega \leq \pi$. The plot shows a slightly smoothed version with a sliding rectangular window of length 33. Notice that this is a log-log plot and the variations near zero-frequency can be seen clearly. The $1/f$ behavior continues till very low frequencies, flattening out only as we get really close to zero frequency. Notice also the thin line representing a sharp peak near the right edge of the plot. This corresponds to the peak at $2\pi/3$ due to period-3 property in the coding regions. More examples can be found in [36]. Li has written a comprehensive review paper on this topic [33], and has also observed [32] that the $1/f$ behavior in natural phenomena can be traced to the so-called duplication-mutation model (see Fig. 18).

[5]Recall that the sample spacing for indicator sequences is normalized to be unity, so the highest frequency π corresponds to 0.5.

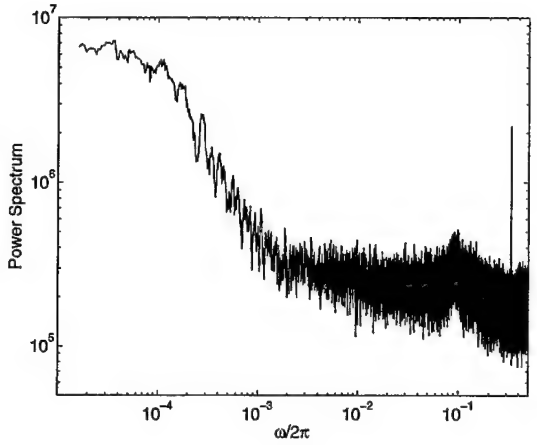


Figure 17. Demonstration of $1/f$ spectrum. The $1/f$ behavior extends to very small frequencies indicating very long range correlation.

In addition to the overall $1/f$ behavior of DNA sequences, and the period-3 property in protein coding regions, it has been observed by many authors that DNA molecules also have components of period 10 to 11 (see [31] and references therein). In [31] it is argued that this periodicity can be attributed to an alternation property in protein molecules. This arises from the fact that the hydrophilic and hydophobic regions (water loving and water hating regions) alternate at a certain rate in the three-dimensional folded form.

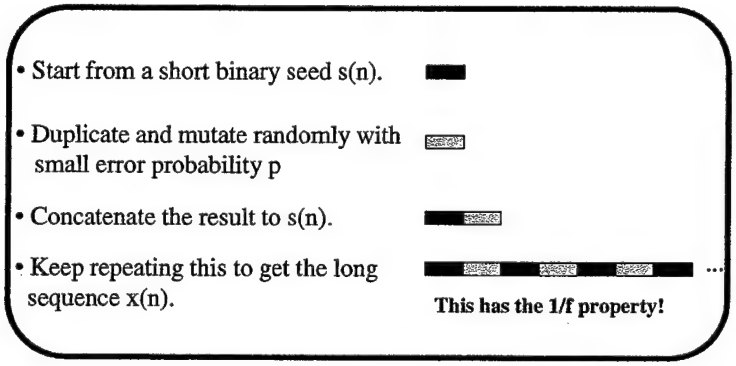


Figure 18. When life started on earth the DNA molecules were short (few thousand bases). As evolution progressed, the molecules went through a lengthening processes which involved duplication and mutation. Imagine we have a character string $s(n)$ of length L. Suppose we duplicate it and then make some random changes of certain characters (from the same alphabet), and concatenate the result to the original $s(n)$ to form a sequence that is twice as long. Further repetition of duplication and mutation quickly results in a very long sequence comparable to today's DNA molecules. It can be shown that repeated application of the duplication-mutation process results in $1/f$ behavior in the spectrum [32].

6.3. Fourier Transforms of Protein Sequences

Of fundamental importance to protein functioning is the ability of a protein to interact selectively with other molecules. This ability comes from the very sophisticated 3D shape assumed by a protein depending on its

amino acid sequence (e.g., Fig. 12, Sec. 5). There are specific sites in the 3D structure called **hot spots** where certain other molecules can conveniently bind to the protein (see the cartoon demonstration in Fig. 19).

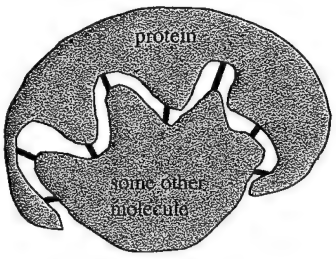


Figure 19. Toy cartoon, showing how the surfaces of certain protein molecules fit like puzzle pieces when they interact.

A protein molecule typically has many functions (many hot spots). Given a collection of proteins, suppose they all have one function in common. Is there a mathematical way to identify this commonality simply by analyzing the amino acid sequence? Yes indeed, based on Fourier techniques [12].

With each one of the twenty amino acids it is possible to associate a unique nonnegative number called the average *electron-ion interaction potential* (EIIP). The physical basis for this is explained in [12] and references therein. The EIIP values are shown in Fig. 20 and plotted in increasing order in Fig. 21. Given a protein, we can associate a numerical sequence $x(n)$ with it such that $x(n)$ is equal to the EIIP value of the nth amino acid. The argument n can be regarded as equispaced distance (≈ 3.8 Å or 0.38 nm, the spacing between amino acids).

Let $X(e^{j\omega}) = \sum_{n=0}^{N-1} x(n)e^{-j\omega n}$ be the Fourier transform of $x(n)$, where N is the number of the amino acids in the protein. Usually a plot of $|X(e^{j\omega})|$ does not reveal much (e.g., see top plots in Fig. 22), Now assume that we have a **group** of proteins. Each protein may have several biological functions but assume that there are some functions that are common to all these proteins. Define the magnitude of the product of the Fourier transforms associated with these proteins as follows: $P(e^{j\omega}) = |X_1(e^{j\omega})X_2(e^{j\omega})\ldots X_M(e^{j\omega})|$. It has been observed through extensive experiments that if a group of proteins has only one common function then the product spectrum $P(e^{j\omega})$ has one significant peak (bottom plot, Fig. 22). This corresponds to the statement that there are common periodic components in the EIIP sequence in the amino acid domain. The physical basis for this arises from the so-called *resonant recognition* between proteins and their targets [12]. The product $P(e^{j\omega})$ has been referred to as the **consensus spectrum** among the group of proteins used in its definition. The frequency where the peak occurs is called the characteristic frequency for the particular protein group. For example the charactereristic frequency is 0.0234 for hemoglobins and 0.3203 for glucagons

(frequencies are normalized to be in the range $[0, 0.5]$ as in standard DSP practice).[6]

Assume we have identified that a certain function of a protein is associated with the characteristic frequency f_1. Is it possible to identify the amino acids that are primarily responsible for that function (i.e., identify the hot spots in the 3D protein structure which are responsible for one particular function)? This is tricky because the value of a Fourier transform at a given frequency depends on all the time-domain samples. Transforms which offer a local basis such as the wavelet transformation and short time Fourier transformation are more convenient and have been successfully used for this [17], [18]. A detailed study of the use of wavelet transforms in protein structures can be found in the recent paper by Murray, et al. [16]. The impact of the use of signal processing tools here could be significant. One advantage of being able to identify a characteristic frequency with a particular funtionality is that it is then possible to synthesize artificial amino acid sequences or peptides (short amino acid sequences). These could be potentially useful in medicine [17].

1	*A*	Ala	Alanine	0.0373	11	*M*	Met	Methionine	0.0823
2	*C*	Cys	Cysteine	0.0829	12	*N*	Asn	Asparagine	0.0036
3	*D*	Asp	Aspartic acid	0.1263	13	*P*	Pro	Proline	0.0198
4	*E*	Glu	Glutamic acid	0.0058	14	*Q*	Gln	Glutamine	0.0761
5	*F*	Phe	Phenylalanine	0.0946	15	*R*	Arg	Arginine	0.0959
6	*G*	Gly	Glycine	0.0050	16	*S*	Ser	Serine	0.0829
7	*H*	His	Histidine	0.0242	17	*T*	Thr	Threonine	0.0941
8	*I*	Ile	Isoleucine	0.0000	18	*V*	Val	Valine	0.0057
9	*K*	Lys	Lysine	0.0371	19	*W*	Trp	Tryptophan	0.0548
10	*L*	Leu	Leucine	0.0000	20	*Y*	Tyr	Tyrosine	0.0516

Figure 20. Electron-ion interaction potentials (EIIP) value for the twenty amino acids [12].

[6]Hemoglobins are oxygen carriers in the red blood cells. Glucagons are protein hormones generated in the pancreas, and affect glucose level in blood.

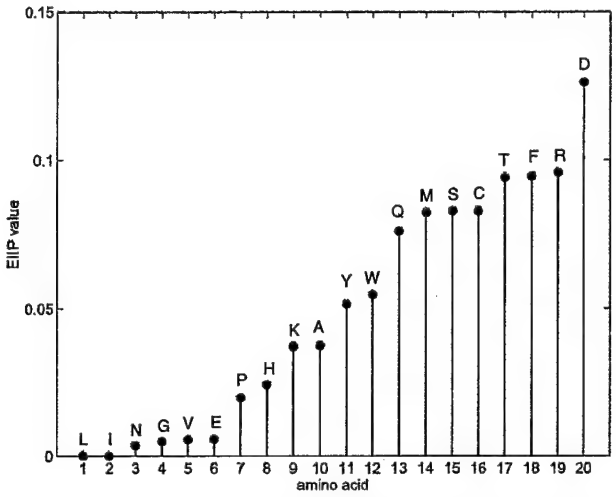


Figure 21. Plot of the electron-ion interaction potential (EIIP) for the twenty amino acids.

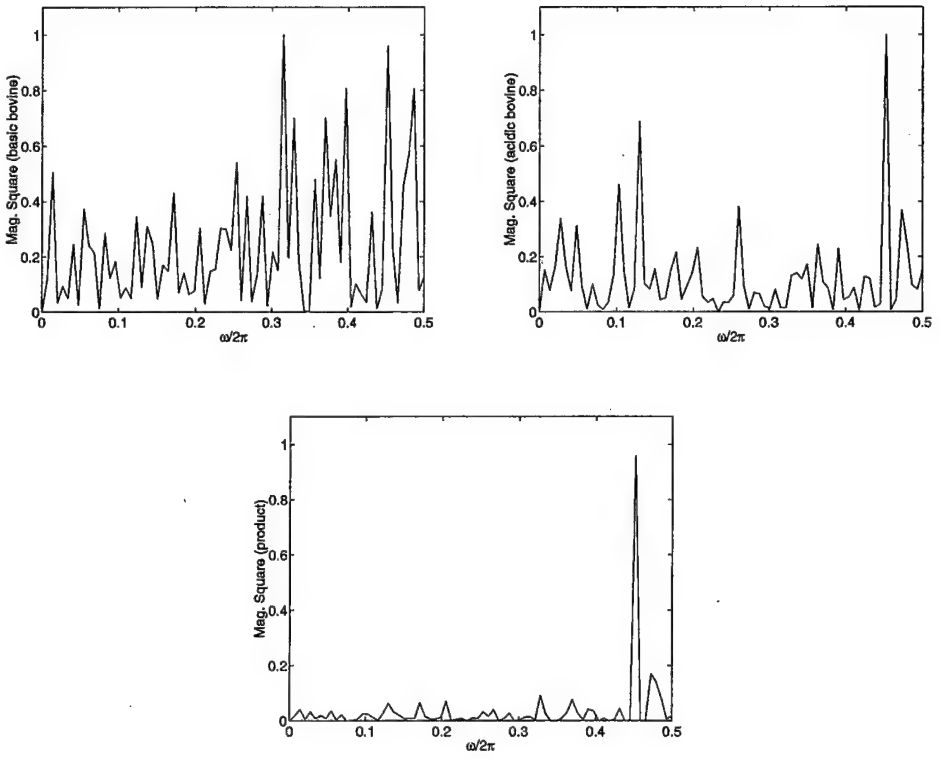
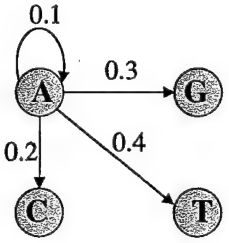


Figure 22. Magnitude squares of the Fourier transforms of the EIIP sequences for the proteins FGF basic bovine (top left) and FGF acidic bovine (top right). The product, which represents the square of the consensus spectrum, is plotted in the bottom [12].

7. ROLE OF HIDDEN MARKOV MODELS

Markov models are very useful to represent families of sequences with certain specific properties. To explain the idea consider Fig. 23(a) which shows part of a DNA sequence. The base A appears a few times, and it can be followed by an A, C, T, or a G. Given a long DNA sequence we can count the number of times the base A is followed by, say, a G. From this we can estimate the probability that an A is followed by a G. If this probability is 0.3 for example, we indicate it as shown in Fig. 23(b). The figure also shows examples of probabilities for A to transition to other bases, including itself. The first row of the matrix in Fig. 23(c) shows the four probabilities more compactly (notice that their sum is unity). Similarly the probabilities that the base C would transition into the four bases can be estimated, and is shown in the second row of the matrix. This 4×4 matrix is called a **state transition matrix**, and is denoted as $\mathbf{\Sigma}$. Fig. 23(b) is called a Markov model. The four **states** in this model are A, C, T, and G. Given a sequence or a set of sequences of "similar kind" (e.g., a long list of exons from several genes) the parameters of the model (the transition probabilities) can readily be estimated. The process of identifying the model parametes is called *training* the model. In all discussions it is implicitly assumed that the probabilities of transitions are fixed and do not depend on past transitions.

(a) **AACTGAGGTACAATTCGATCTC**



(b)

(c)

	A	C	T	G
A	0.1	0.2	0.4	0.3
C	0.2	0.5	0.1	0.2
T	0.5	0.2	0.1	0.2
G	0.3	0.1	0.4	0.2

State transition matrix $\mathbf{\Sigma}$

Figure 23. Explaining the basic principle of the Markov model. (a) A sequence of bases, (b) the state diagram showing the transitions from A, and (c) an example of the state transition matrix.

Suppose we are given a Markov model (i.e., $\mathbf{\Sigma}$ given). Given an arbitrary state sequence $\mathbf{x} = [x(1), x(2), \ldots, x(L)]$ we can calculate the probability that $\mathbf{x}$ has been generated by our model. This is given by the product

$$P(\mathbf{x}) = P(x(1)) \times P(x(1) \rightarrow x(2)) \times P(x(2) \rightarrow x(3)) \times \ \ldots \ \times P(x(L-1) \rightarrow x(L))$$

where $P(x(k) \rightarrow x(m))$ is the transition probability for going from $x(k)$ to $x(m)$, and can be found from

the matrix $\mathbf{\Sigma}$. The usefulness of such computation is as follows: given a number of Markov models ($\mathbf{\Sigma}_1$ for introns, $\mathbf{\Sigma}_2$ for exons, and so forth) and given a sequence $\mathbf{x}$, we can calculate the probabilities that this sequence is generated by any of these models. The model which gives the highest probability is most likely the model which generated the sequence.

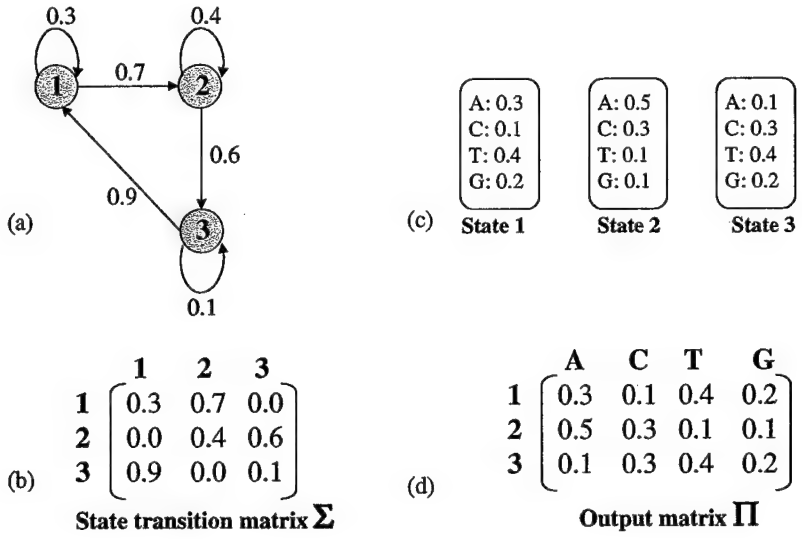


Figure 24. Basic principle of the hidden Markov model (HMM). (a) State diagram, (b) state transition matrix, (c) state to output probabilities, and (d) output matrix.

A hidden Markov model (HMM) is obtained by a slight modification of the Markov model. Thus consider the state diagram shown in Fig. 24(a) which shows three states numbered 1, 2, and 3. The probabilities of transitions from the states are also indicated, resulting in the state transition matrix $\mathbf{\Sigma}$ shown in Fig. 24(b). When the system is in a particular state, it can output one of four possible symbols, namely A, T, C, or G, and there is a probability associated with each of these. This is demonstrated in Fig. 24(c), and summarized more compactly in the so-called **output matrix $\mathbf{\Pi}$** shown in Fig. 24(d). To give an example of how HMMs might be useful, we can imagine that state 1 corresponds to exons, state 2 to introns, and state 3 to intergenic spaces. In each of these states, the probabilities of transitions between bases could be different.

In order to apply the hidden Markov model theory successfully there are three problems that need to be solved in practice [6]. These are listed below along with names of standard algorithms which have been developed for these.

1. Given an HMM (i.e., given the matrices $\mathbf{\Sigma}$ and $\mathbf{\Pi}$) and an output sequence $y(1), y(2), \ldots$, compute the state sequence $x(k)$ which most likely generated it. This is solved by the famous **Viterbi's algorithm**.

2. Given the HMM and an output sequence $y(1), y(2), \ldots$, compute the probability that the HMM generates this. The **forward-backward algorithm** solves this.

3. The third problem is training: how should one design the model parameters $\mathbf{\Sigma}$ and $\mathbf{\Pi}$ such that they are optimal for an application, e.g., to represent exons? The most popular algorithm for this is the expectation maximization algorithm commonly known as the **EM** algorithm or the **Baum-Welch** algorithm.

Further details on these algorithm can be found in [6]. The theory of HMMs has been applied successfully for gene identification, for identification of special regions of DNA such as CpG islands, and for DNA sequence alignment. There are many good references which explain the use of HMMs in molecular biology. A good start would be to look at [24], [25], [7], and [4], and then proceed to references therein. As for basics, there are excellent tutorials and books which explain the theory of Hidden Markov models. The paper by Rabiner in the Proceedings of the IEEE [6] has been widely cited in the molecular biology literature. The books by Rabiner and Juang [66] and by Jelinek [61] give wonderful exposure to the theory and its applications in speech recognition.

8. NON CODING GENES AND ncRNA

The most common meaning associated with genes during the four decades following the discovery of the double helix was that genes are those parts of the DNA sequence that code for proteins (Sec. 3). But it has become increasingly clear in the last ten years that there are portions of DNA which are transcripted to RNA sequences that *do not get translated to proteins.* These are called noncoding RNA or **ncRNA**, and the portions of DNA which generate them are called **noncoding genes**. Many of these are located in the intergenic space (space between protein coding genes). Indeed ncRNAs have been known for many years, the transfer RNA (tRNA) and ribosomal RNA (rRNA) being classic text-book examples [2]. However, the recognition that there are many different ncRNAs and that noncoding genes play a heriditary role is more recent. The fact that noncoding genes have such tremendous importance has been regarded as a challenge to the central dogma of molecular biology which suggests that genes by definition code for proteins (Sec. 3). So the intergenic space cannot by any means be regarded as "junk DNA" as it used be to at one time. An excellent place to start reading about noncoding genes is the Scientific American article by Gibbs [42]. Papers by Eddy such as the Nature genetics review article [40] are informative as well as insightful.

Figure 25. Left: the C. elegans worm, mangified many times. Right: if a human baby grew in size but not in features, that would be analogous to the C. elegans story which lead to the discovery of the importance of ncRNA genes. See text.

Perhaps the discovery of the importance of noncoding genes can be traced back to the case of a C. elegans baby that wouldn't grow up.[7] In an observation made by Ambrose et al. (Dartmouth medical school, Hanover, N. H) there was a C. elegans baby in the laboratory which grew in size but never came out of the first larva stage (see analogy of human baby, Fig. 25). The scientists were able to trace this to a defective gene. In the healthy worm the function of this gene was to produce a tiny RNA molecule, only 22 bases long. The role of this RNA molecule was to regulate other protein coding genes responsible for normal growth into adult. So this RNA did not get translated into a protein; it was an ncRNA, and functioned all by itself. In the defective C. elegans baby, this particular ncRNA gene in the DNA was mutated, and the ncRNA was not functioning properly, thereby affecting growth functions. This was the first ncRNA recognized (besides tRNA and so forth), and ncRNAs were taken seriously only after this observation. See the short but fascinating account given by John Travis in [49].

Many more ncRNAs have been found in several organisms in the last ten years and their functions identified [40], [48]. It has been conjectured [42] that about fifty percent of the genes in mice generate ncRNAs rather than proteins! C. elegans has more than 200 genes generating micro ncRNAs (tiny ncRNAs about 22 bases long). And the E. coli bacterium has several hundred noncoding genes and about 4200 protein coding genes [40]. Today it is recognized that heriditary information is carried by protein coding genes, noncoding genes and a third layer of information storage called the epigenetic layer [43].

Noncoding genes have created a great deal of excitement in medicine. Other related research not discussed here include the role of double strand RNAs and antisense RNAs in gene silencing. These are called siRNAs (small interfering RNAs) and can be inserted into cells to prevent the expression of hazardous genes. A good starting point for the interested reader is the series of Scientific American articles [43]–[45]. The discovery of noncoding genes apparently solves a long-held puzzle in biology. It has been known that the number of protein coding genes never scales in proportion to the size of the organism [42]. For example worms have

[7] C. Elegans is a worm or nematode used extensively in biological studies. It grows into an adult with exactly 959 cells.

only twice as many protein-coding genes as bacteria. Humans have only thrice as much (about 35,000). And the rice plant has more genes than humans! But if the number of noncoding genes are counted, it seems that the total number of genes does scale well with complexity of organisms [42].

8.1. Identifying noncoding genes

In Sec. 6.1 we explained that there are many ways to identify protein coding genes in DNA sequences. These genes have a period-3 component due to codon bias which is usually quite strong. For more precise identification one can use hidden Markov models as explained in [24]. Computational identification of noncoding genes is much more difficult. These genes could be very small (sometimes no longer than 22 bases), do not exhibit the period-3 property, and do not have start and stop codons. Conclusions could often be wrong; there are case histories where certain genes, originally thought to be ncRNA genes, were later found to encode tiny proteins [40].

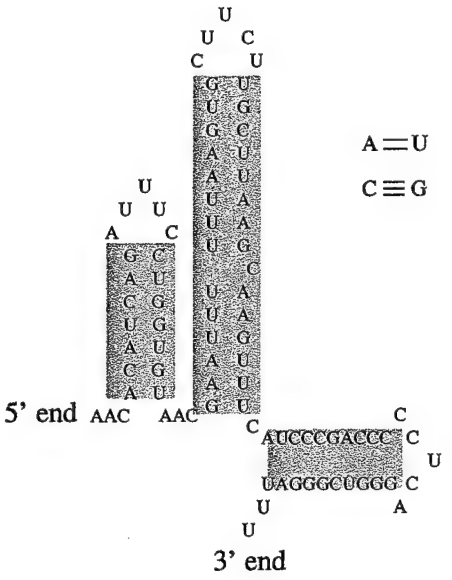


Figure 26. The picture shows a dsrA RNA in E. coli. Notice the secondary structure created by base pairing in blue shaded areas. Functionality of ncRNAs depends mostly on their secondary structure.

It has been noticed that noncoding genes and ncRNAs function by virtue of their **secondary structure** which we explain next. Consider Fig. 26 which shows an ncRNA in E. coli bacteria. Notice that even though it is a single stranded molecule (like most RNAs are), there are long stretches of bases in one part which are complementary to stretches in other parts (recall here that A pairs up with U and C with G). This complementarity forces the RNA to fold into shapes which are not only beautiful, but in fact are crucial to

their biological functioning. Many of the RNAs can act as enzymes primarily by virtue of this folded shape. RNA enzymes are called ribozymes, so they are not confused with normal enzymes which are proteins. Some computational biologists have suggested that noncoding genes in the DNA sequences can be identified simply by looking for subsequences which have secondary structure [47].

We will return to this later but briefly mention another approach called **comparative genomics** which has been reasonably successful. The idea behind comparative genomics is that if two or more species have a common stretch of DNA, then it is probably doing something important. Otherwise nature would not have conserved it for millions of years. So these stretches would have to be either protein coding genes or noncoding genes. If they do not pass standard tests for protein coding genes they are likely to be noncoding genes. In this way it is possible to accumulate a list of potential noncoding genes in a given species and then check them by other biological means. Comparative study of DNA sequences is not as simple as it appears to be on first sight because the sequences being compared come from various species, and "identical regions" can still differ due to mutations, insertions, and deletions of bases through millions of years of evolution. For example consider the following four sequences:

$$\times\times\times AATAGCGA\times\times\times\times\times\times\times\times\times\times\times AATAC\times\times\times AAATACCG$$

$$\times\times\times\times\times\times\times AATAGCGA\times\times\times\times\times AATAC\times\times\times\times\times AAATACCG$$

$$\times\times\times\times\times\times\times AAGAGCGA\times\times\times\times\times AATAC\times\times\times\times\times AAAGTCCG$$

$$\times\times\times\times\times\times\times AAAGCGA\times\times\times\times\times AATAC\times\times\times\times\times AAATAAACCG$$

where $\times$ denotes that the base could be any one of the four. Inspection reveals that there are many common patterns here. However a direct comparison base by base would lead a computer to conclude that these are not identical sequences at all. There are patterns which are common but with slight mutations; there are unequal gaps between similar patterns; and the "identical parts" often do not even have identical lengths!

The task of comparing such sequences is nontrivial science. It comes under the topic of **sequence alignment**. Computational biologists have developed many methods for this and in fact assign scores to degree of similarity between sequences. Markov models have been used for this application. Many wonderful details can be found in the book by Durbin et al. [4]. In a study by the National Human Genome Research Institute (NHGRI) the human genome has been compared with many others such as cows, dogs, pigs, and rats. It has been found that there were over 150 common regions in the intergenic space! Many potential ncRNA sequences have been listed in this way and later confirmed by other means. The method of comparative genomics to identify ncRNAs does not work perfectly yet, but has been quite useful.

8.2. Identifying Secondary Structure

A few words on identification of secondary structures directly without comparative genomics. Consider Fig. 27(a) which shows a DNA sequence with two short subsequences AATC and GATT buried in it. These subsequences are separated by many bases. If we reverse the first subsequence we get CTAA which is complementary to the second sequence. So the subsequences can be regarded as two halves of a palindrome (i.e., a symmetric sequence like $xyzpqpzyx$).[8] The sequence can therefore fold as shown in Fig. 27(b) and remain stable in that configuration because of the A—T and C—G bondings. If an ncRNA is generated from such a DNA segment it would therefore fold as shown. In practice the matching subsequences many not match exactly, they may be separated by arbitrary number of bases, and furthermore there may be more than one matching pair. The secondary structure can therefore be quite complicated. All of these features can be clearly seen in the example of ncRNA shown in Fig. 26.

Figure 27. (a) Example of a palindrome-like pattern buried in DNA, and (b) the natural way for this sequence to fold.

The biological functioning of the ncRNA depends primarily on the way it folds, that is, on the secondary structure rather than the exact sequence of base pairs. For example the two sequences shown below would fold the same way.

$$\times\times\times AATC\times\times\times\times\times\times\times\times\times\times\times GATT\times\times\times$$

$$\times\times\times AGTA\times\times\times\times\times\times\times\times\times\times\times TACT\times\times\times$$

Computational identification of ncRNA genes is therefore closely related to the identification of buried pattens such as palindromes in a long arbitrary sequence (a few thousand or million bases). This is quite a challenging problem. One of the theoretical bottlenecks is that hidden Markov models which worked so well for identification of protein coding genes do not work anymore as explained next.

[8]Not exactly a palindrome because of the complement operation, but we shall refrain from inventing a new word for that.

8.3. Grammars

In the language of computer science, a grammar is a set of rules which can be repeatedly applied to obtain sequences of letters from an alphabet. The set of all sequences that can be generated by a grammar is called the language generated by that grammar. In the early 1950s, Noam Chomsky (a phenomenal computational linguist from MIT) classified grammars into four types called regular grammars, context free grammars, context sensitive grammars, and unrestricted grammars. The relation between these grammars is depicted in Fig. 28.

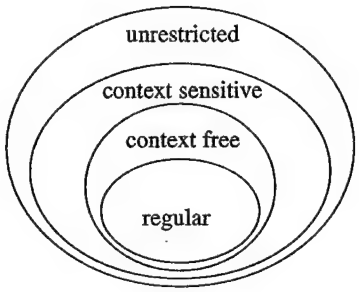


Figure 28. Chomsky's hierarchy of grammars for generating languages.

Regular grammars have the most restricted production rules and therefore generated a restricted class of languages. Context free grammars allow a wider class of production rules and generate a broader class of languages. For example suppose the "language" is the set of all palindromes. Then there is no regular grammar to generate these, but there does exist a context free grammar.[9]

We now give a very brief overview of grammars. Good references to this topic include [60] and [62]. A **regular grammar** allows production rules of the form $W \to aW$ and $W \to a$, where W is a **nonterminal** symbol (i.e., we can substitute further for it) and a is a **terminal** symbol. Consider the example of a regular grammar with the following three production rules, where A, C, and T are the terminals:

$$W \to AW, \quad W \to TW \quad W \to CW, \quad W \to A, \quad W \to T, \quad \text{and} \quad W \to C.$$

Here is an example of a string generated by this grammar by repeated application of the rules in arbitary order:

$$W \to AW \to AAW \to AACW \to AACTW \to AACTT$$

The language generated by this grammar is the string of all DNA sequences with the base G missing.

[9]True, we can find a regular grammar which generates palindromes among other possible sequences. But we cannot find a regular grammar which generates only palindromes.

A **context free grammar** allows production rules of the form $W \rightarrow \alpha$ where W is a nonterminal and α is a string of terminals and nonterminals. A grammar defined by the following production rules is an example. Here A, C, G, and T are the terminals.

$$W \rightarrow AWA, \quad W \rightarrow CWC, \quad W \rightarrow TWT \quad W \rightarrow GWG \quad \text{and} \quad W \rightarrow \epsilon$$

where ϵ represents the null string (i.e., nothing). Here is an example of a string generated by this grammar:

$$W \rightarrow AWA \rightarrow ATWTA \rightarrow ATCWCTA \rightarrow ATCCTA$$

In the last step W has been replaced with the null terminal character. Notice that the resulting string is a palindrome. The preceding grammar generates the palindromes language.

If the production rules in a grammar are used with a certain probability attached to each rule, it is called a **stochastic grammar**. There is a result in the theory of computations which says that *stochastic regular grammars are identical to hidden Markov models.* That is, if a class of strings can be generated by a stochastic regular grammar then there exists an HMM which generates this class, and vice versa. Since regular grammars cannot generate palindrome languages we cannot therefore build HMMs that represent noncoding genes. We cannot therefore use HMM theory to identify noncodig genes buried in long DNA sequences. Stochastic context free grammars, abbreviated as **SCFG**s, have been used for this purpose and a great deal of detail can be found in [4] and references therein. Figure 29 summarizes some of these discussions.

Recall from Sec. 7 that in order to apply the HMM theory successfully there are three problems that need to be solved, and there exist standard algorithms for this, namely Viterbi's algorithm, forward-backward algorithm, and the EM algorithm. For the case of context free grammars there are similar algorithms but they have much higher complexity [4]. The importance of fast procedures for these arises because of the fact that DNA sequences are very long even for "small" organisms. Computational biologists are therefore interested in developing faster algorithms for the above problems. Recently Yoon and Vaidyanathan have introduce a class of hidden Markov models called **context sensitive HMMs** [50] which appear to be promising for this application while at the same time offering significantly lower complexity.

Finally, even context sensitive languages have had some applications in this context. An example of a language that can be recognized by such grammars but not by context free grammars is the so-called copy language [4] which can sometimes be useful in describing secondary structures.

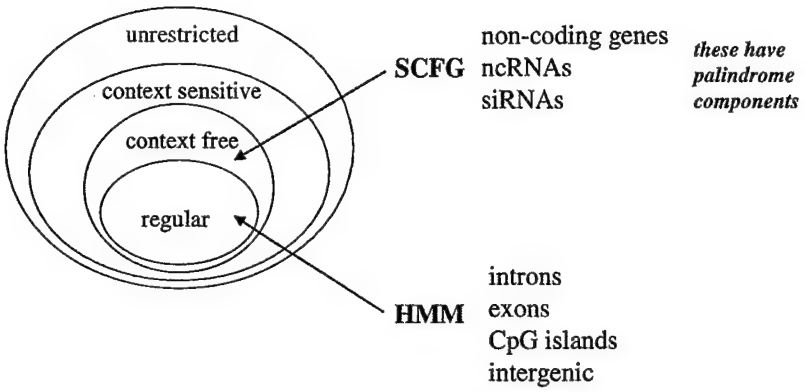


Figure 29. Application of grammars in computational biology. Regular grammars (stochastic hidden Markov models) are useful for identifying protein coding genes whereas stochastic context free grammars (SCFGs) are necessary to identify noncoding genes.

9. OTHER AREAS

In the past few sections a number of interesting areas were discussed but many were also left out for want of space. One of these is **DNA computation**. The enormous capabilities of the cell (base-pairing, gene-protein feedback) can be used to perform miraculously difficult computational tasks. A starting point for the reader would be the article by Adleman in 1998 in the Scientific American [59]. Another area we did not discuss is DNA sequencing. Many signal processing aspects are involved here, and a flavor can be obtained by reading [14] and [22]. An informal discussion of some other areas is given here with appropriate pointers to literature.

9.1. DNA Microarrays

An entire issue of Nature genetics was dedicated to the topic of DNA microarrays in 1999. The reader should see [53] and other articles therein for an excellent introduction. A good overview also appeared in the IEEE Spctrum a few years ago [57], so we will be brief. DNA microarrays are typically grown on a piece of glass or silicon substrate chemically primed so that the molecules A, C, T and G stick to specific sites. It is possible to raise towers of base sequences about 100 bases long, using photolithography as shown in Fig. 30. In this way an entire gene can be "grown" on a few towers. Several genes can therefore be captured onto a single DNA microarray chip.

These chips can be used to observe the expression levels of different genes in the cell as explained in Fig. 31. The real advantage here is that we can measure the levels of several genes simultaneously, and as a function of time (e.g., cell cycle) and so forth. This gives an enormous advantage to biologists who wish to study the dependency of gene expressions on various factors. An example is the 1999 experiment at MIT [57] where Affymetrix chips containing 6800 human genes were used to analyze the expression of genes in

cancer cells from two types of blood cancer (acute myeloid leukemia and lymphoblastic leukemia). Standard pathology examination failed to distinguish the two types but the arrays showed a set of 50 genes that have different activity levels in the two cancers. Many examples can be found in the papers published in Nature genetics, Jan. 1999, and papers such as [51]. DNA microarrays have serious application in drug design [55], antiterrorism [54], and many other related areas.

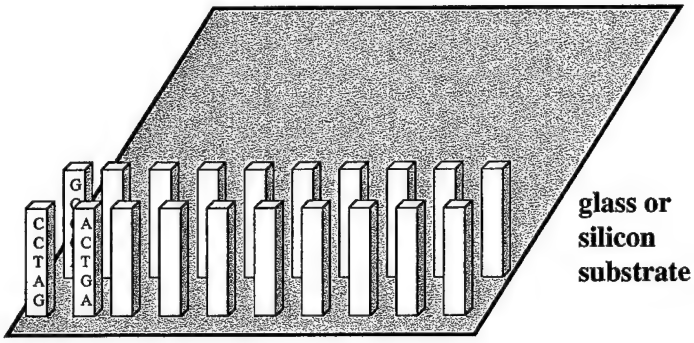


Figure 30. The DNA microarray.

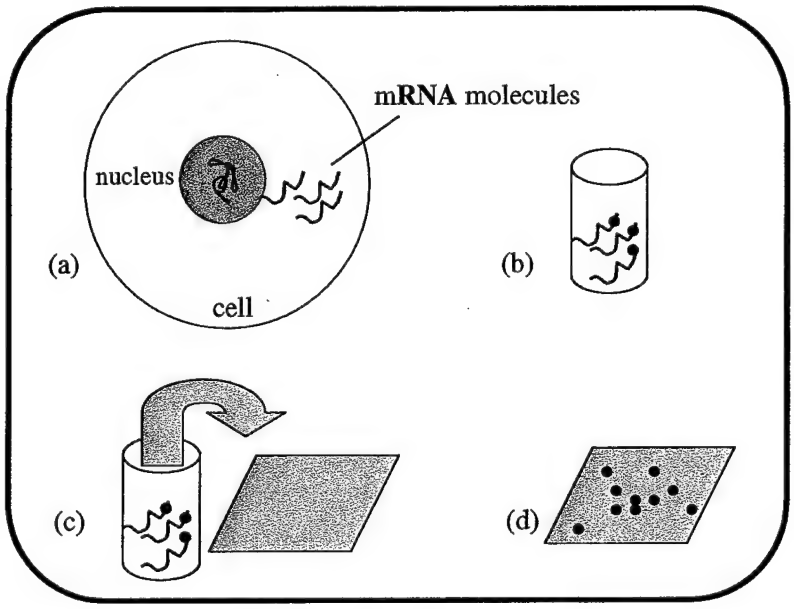


Figure 31. Measuring gene expression levels using DNA microarrays. (a) When a gene is expressed, mRNA molecules are released from the nucleus. (b) These mRNA molecules are collected, turned into DNA, and tagged with florescent dyes. (c) This gene-cocktail is poured onto the gene micro-array. The tagged molecules stick to the portions of genes on the array which are *complementary* to them. (d) If the array is now illuminated with a light source, then the tags glow and one can see which genes are expressed. Strength of the glow indicates the amount of gene expression.

The Affymetrix series started with a modest 1000 genes on chip in 1998. Today, nearly all of the protein

coding genes in humans (about 35,000) have gone into a single chip (Affymetrix Inc. and Agilent Technologies announced these in 2003). There are some interesting signal processing issues involved in the interpretation of data recorded on a DNA array. Some examples can be found in [58] and [51].

9.2. The Gene-Protein feedback loop

We know genes guide the generation of proteins. But proteins to a large extent also control which genes are expressed and to what extent. In short, proteins can switch genes on and off. The gene-protein feedback loop is what make different cells look and function differently. Cell function depends on a gene-protein network interconnected in a highly complex manner.

The first hint that proteins in cells might be influencing gene expression came from Francois Jacob and Jacques Monad in Paris, around 1960. The E. coli bacteria uses lactose sugar and breaks it into simpler sugars (galactose and glucose) using the enzyme *beta galactosidase*. When lactose is absent in the bacterial medium the E. coli cell does not produce this enzyme. Otherwise it does! Jacob and Monad suggested that this switching ability is due to the presence of a repressor molecule. In the late 60s Walter Gilbert and Benno Müller-Hill (from Harvard) found the molecule. The repressors were proteins and this was the first proof that there is a closed loop (feedback) system. In recent years, the closed loop relation has been described with some success using linear first order coupled differential equations called Langevin equations [19], and this has been found to be useful in systematic analysis of uncertainties (or "noise") in gene circuits. A fascinating account of information processing in genetic circuits can be found in the May 2004 IEEE paper by Simpson, et al. [20].

9.3. Relation to RNA world

If proteins are generated by genes and genes are in turn controlled by proteins, then which came first? This is similar to asking whether the chicken or egg came first. The fact that ncRNA molecules can perform many of the functions of proteins (Sec. 8) answers this question to some extent. There is a theory called the *RNA-world* theory which suggests that the earliest form of life on earth was based entirely on RNA molecules. Some of these RNA molecules carry genetic information (like genes in DNA), wheras some act as catalysts.[10] The article by Orgel in the Scientific American [46] traces the origin of this theory and gives an account of some laboratory experiments which demonstrate the feasibility of the RNA-world theory.

10. CONCLUDING REMARKS

In this article we have attempted to share the excitement of molecular biology from the point of view of the scientist with a signal processing and circuits background. We conclude with the sentiment that genomics,

[10]RNA catalysts are called ribozymes rather than enzymes; the latter name is reserved for protein catalysts.

and more generally molecular biology have taken a very interesting turn for all of us. For those who did not like biology because of the wet smelly labs, there is good news. Molecular biology today involves signal processing, computer science, mathematics, and informatics, all coming together beautifully!

Acknowledgements. It is my pleasure to thank Profesors A. Antoniou, L. Bruton, and M. N. S. Swamy, for strongly encouraging me to write this article. I am grateful to Dr. M. Ogorzalek for all his support and help during the production of this article. Many of the Fourier transform and digital filtering plots in this paper were prepared by Mr. Byung-Jun Yoon at the California Institute of Technology.

REFERENCES

We have tried to categorize the papers into subtopics. Many papers can easily belong in more than one category. So please do not overlook any of these. The selection here is by no means complete but based on my personal taste. Perhaps a good list to start with, to teach from, to make reading-assignments from, and so forth.

The paper which started it all ...

[1] J. D. Watson, and F. H. C. Crick, A structure for DNA, *Nature*, p. 737, April 1953.

Books and Tutorials

[2] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology*, Garland Publishing Inc., New York, 1998.

[3] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, pp. 8–20, July 2001.

[4] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge Univ. Press, 1998.

[5] Proc. of IEEE special issues Dec. 2000 (Genomic Engineering), Nov. 2002 (Bioinformatics, part 1: advances and challenges), and Dec. 2002 (Bioinformatics, part 2: genomics and proteomics engineering).

[6] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. of the IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[7] S. L. Salzberg, D. B. Searls, and S. Kasif, *Computational methods in molecular biology*, Elsevier, 1998.

[8] P. P. Vaidyanathan, and B-J. Yoon, "The role of signal processing concepts in genomics and proteomics," Journal of the Franklin Institute, vol. 341, pp. 111–135, 2004.

[9] J. D. Watson, *The double helix*, Simon and Schuster, N. Y., 1968.

[10] J. D. Watson, *Genes, girls, and Gamow*, Vintage books, Random House, Inc., N. Y., 2001.

[11] J. D. Watson (with Andrew Berry) *DNA: the secret of life*, Alfred A. Knopf, N. Y., 2003.

Signal-processing flavor (DNA/Protein)

[12] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules? — theory and applications", IEEE. Trans. Biomedical Engr., vol. 41, no. 12, pp. 1101–1114, Dec. 1994.

[13] S. W. Davies, M. Eizenman, and S. Pasupathy, "Optimal structure for automatic processing of DNA sequences," IEEE Trans. on Biomedical Engr., vol. 46, no. 9, pp. 1044–1056, Sept. 1999.

[14] W. Huang, D. R. Fuhrmann, D. G. Politte, L. J. Thomas, and D. J. States, "Filter matrix estimation in automated DNA sequencing," IEEE Trans. on Biomedical Engr., vol. 45, no. 4, pp. 422–428, April 1998.

[15] Koradi, R., Billeter, M., and Wthrich, K., "MOLMOL: a program for display and analysis of macromolecular structures', J. of Mol. Graphics vol. 14, pp. 51-55, 1996.

[16] K. B. Murray, D. Gorse, and J. M. Thornton, "Wavelet transforms for the characterization and detection of repeating motifs," J. Molecular Biology, vol. 316, pp. 341–363, 2002.

[17] E. Pirogova, Q. Fang, M. Akay, and I. Cosic, "Investigation of the structural and functional relationships of oncogene proteins", Proc. of the IEEE, vol. 90, no. 12, pp. 1859–1867, Dec. 2002.

[18] P. Ramachandran, A. Antoniou, and P. P. Vaidyanathan, "Identification and location of hot spots in proteins using the short-time Fourier transform" IEEE Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, Nov. 2004.

[19] M. L. Simpson, C. D. Cox, and G. S. Sayler, "Frequency domain analysis of noise in autoregulated gene circuits" Proc. of the Nat. Academy of Sci., vol. 100, no. 8, pp. 4551-4556, April 15, 2003.

[20] M. L. Simpson, C. D. Cox, G. D. Peterson, and Gary S. Sayler, "Engineering in the biological substrate: information processing in genetic circuits," Proc. of the IEEE, vol. 92, no. 5, pp. 848–863, May 2004.

[21] D. Sussillo, A. Kundaje, and D. Anastassiou, "Spectrogram analysis of genomes", Eurasip J. of Applied Signal Processing, vol. 2003, no. 4, Dec. 2003.

[22] X-P. Zhang, and D. Allison, "Iterative deconvolution for automatic base scaling of the DNA electrophoresis time series," Workshop on Genomic Sig. Proc. and Stat., Raleigh, NC, Oct. 2002.

Gene prediction

[23] J. W. Fickett, "The gene prediction problem: an overview for developers", Computers Chem., vol. 20, no. 1, pp. 103–118. 1996.

[24] A. Krogh, I. Saira Mian, and D. Haussler, "A hidden Markov model that finds genes in E. Coli DNA", Nucleic Acids Research, vol. 22 pp. 4768–4778, 1994.

[25] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models," Nucleic Acids Research, vol. 26, no. 2, pp. 544–548, 1998.

[26] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," CABIOS, vol. 13, no. 3, pp. 263–270, 1997.

[27] P. P. Vaidyanathan, and B-J. Yoon, "Gene and exon prediction using allpass-based filters," Workshop on Genomic Sig. Proc. and Stat., Raleigh, NC, Oct. 2002.

[28] P. P. Vaidyanathan, and B-J. Yoon, "Digital filters for gene prediction applications," IEEE Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, Nov. 2002.

[29] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/.

Long range correlations, $1/f$ behavior, statistics

[30] H. Hausdorff and C.-K. Peng, "Multiscaled randomness: a possible source of $1/f$ noise in biology," Physical review E, vol. 54, no, 2, pp. 2154–2157, August 1996.

[31] H. Herzel, E. N. Trifonov, O. Weiss, and I. Groβe, "Interpreting correlations in biosequences," Physica A, vol. 249, pp. 449–459, 1998.

[32] W. Li, "Expansion-modification systems: A model for spatial $1/f$ spectra", Physical review A, The American Physical Society, vol. 43, no. 10, pp. 5240–5260, May, 1991.

[33] W. Li, "The study of correlation structures of DNA sequences: a critical review", Computers Chem., vol. 21, no. 4, pp. 257–271, 1997.

[34] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, "Long-range correlations in nucleotide sequences," Nature, vol. 356, pp. 168–170, March 1992.

[35] E. N. Trifonov, and J. L. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence", Proc. of the Nat. Acad. Sci., USA, vol. 77, pp. 3816–3820, 1980.

[36] M. de Sousa Vieira, "Statistics of DNA sequences: a low-frequency analysis," Physical Review E, The American Physical Society, vol. 60, no. 5, pp. 5932–5937, Nov. 1999.

[37] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," Physical Review Letters, vol. 68, no. 25, pp. 3805–3808, June 1992.

[38] G. W. Wornell, "A Karhunen-Loeve-like expansion for $1/f$ processes via wavelets," IEEE Trans. on Information Theory, vol. 36, no. 4, pp. 859–861, July 1990.

[39] Z-G. Yu, V. V. Anh, and B. Wang, "Correlation property of length sequences based on global structure of the complete genome", Physical review E, The American Physical Society, vol. 63, pp. 011903-1—011903-8, 2000.

Noncoding genes, ncRNA

[40] S. R. Eddy, "Noncoding RNA genes and the modern RNA world," Nature reviews, GENETICS, vol. 2, pp. 919–929 Dec. 2001.

[41] S. R. Eddy, "Computational genomics of noncoding RNA genes," Cell, vol. 109, pp. 137–140, April 2002.

[42] W. W. Gibbs, *The unseen Genome: gems among junk*, Scientific American, pp. 48-53, Nov. 2003.

[43] W. W. Gibbs, *The unseen Genome: beyond DNA*, Scientific American, pp. 108-113, Dec. 2003.

[44] W. W. Gibbs, *Synthetic life*, Scientific American, pp. 75-81, May 2004.

[45] N. C. Lau and D. P. Bartel, *Censors of the genome,* Scientific American, pp. 35-41, August 2003.

[46] L. E. Orgel, "The origin of life on earth", Scientific American, pp. 77-83, Oct. 1994.

[47] E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy, "Computational identification of noncoding RNAs in E. Coli," Current biology, vol. 11, pp. 1369–1373, Sept. 2001.

[48] G. Storz, "An expanding universe of noncoding RNAs," Science, vol. 296, pp. 1260–1263, May 2002.

[49] Travis, J. "Biological Dark matter," Science News Online, Jan. 12, 2002. (www.sciencenews.org/articles /20020112/bob9.asp).

[50] B-J. Yoon, and P. P. Vaidyanathan, "HMM with auxiliary memory: a new tool for modelling RNA secondary structures", IEEE Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, Nov. 2004.

DNA microarrays

[51] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling", Proc. of the Natl. Acad. of Sci., USA, vol. 97, no. 18, pp. 10101–10106, Aug. 2000.

[52] O. Alter, P. O. Brown, and D. Botstein, "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms," Proc. of the Natl. Acad. of Sci., USA, vol. 100, no. 6, pp. 3351–3356, March 2003.

[53] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," Nature genetics supplement, vol. 21, pp. 33-37, Jan. 1999.

[54] R. Casagrande, *Technology against terror*, Scientific American, pp. 82-87, Oct. 2002.

[55] C. Debouck and P. N. Goodfellow, "DNA microarrays in drug discovery and development," Nature genetics supplement, vol. 21, pp. 48–50, Jan. 1999.

[56] E. S. Lander, "Array of hope," Nature genetics supplement, vol. 21, pp. 3-4 Jan. 1999.

[57] S. K. Moore, "Making chips to probe genes", pp. 54–60, IEEE Spectrum, vol. 38, no. 3, March 2001.

[58] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, "Iterative normalization of cDNA microarray data," IEEE Trans. Info. Tech. in Biomed., vol. 6, no. 1, pp. 29–37, March 2002.

Other related general references...

[59] L. M. Adleman *Computing with DNA*, Scientific American, pp. 54-61, Aug. 1998.

[60] Aho, A. V., Hopcroft, J. E., and Ullman, J. D., *The design and analysis of computer algorithms,* Addison Wesley Publ. Co., Reading, MA, 1974.

[61] F. Jelinek, *Statistical methods for speech recognition,* The MIT Press, Cambridge, MA, 2001.

[62] H. R. Lewis, and C. H. Papadimitriou, *Elements of the theory of computation*, Prentice Hall, Inc., Englewood Cliffs, N. J., 1981.

[63] Y. Neuvo, and C.-Y. Dong, and S. K. Mitra, "Interpolated finite impulse response filters," IEEE Trans. on ASSP, pp. 563–570, June. 1984.

[64] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing,* Prentice Hall, Inc., NJ, 1999.

[65] A. Papoulis, *Systems and transforms with applications in optics*, McGraw Hill, 1968.

[66] Rabiner, L. R., and Juang, B-H., *Fundamentals of speech recognition,* Prentice Hall, Inc., Englewood Cliffs, N. J., 1993.

[67] P. P. Vaidyanathan, *Multirate systems and filter banks,* Prentice Hall, Inc., NJ, 1993.

Box B1. Story of how the DNA double helix was discovered

The notion that there are specific factors (genes) that are passed on to offspring probably started with the work of Gregor Mendel around 1856. Nearly half a century later it became clear due to the work of Walter Sutton (medical student, Columbia University) and T. H. Morgan (also at Columbia), that these "factors" were located on chromosomes which were known to contain proteins and DNA molecules. In 1930 the DNA was shown to be a long molecule made of the nitrogenous bases A, T, C and G.

In those days proteins were considered to be the "genes" that carried hereditary information. In 1944 it was shown from the experiments of O. T. Avery (Rockfeller Inst., NY) that DNA, rather than protein, carried genetic traits. For example when a virus attacks bacteria, it is the viral DNA and not the protein that enters the bacteria and changes its behavior. Alfred Hershey and Martha Chase verifed this experimentally (1952, Cold Spring Harbor). It was accepted that genes were contained in the DNA; nothing was known about their nature or how they worked.

In 1944, the famous physicist Schrodinger had written a book entitled *What is life*, which inspired many young scientists. J. D. Watson (born in Chicago, 1928) was among them. He was fascinated from childhood about the mystery behind genes. Watson worked on bacteriophages or phages (viruses that attack bacteria) to receive his Ph.D at the young age of 22 from Indiana University, and later went to the Cavendish Laboratories (Cambridge, England) for further work. The following story is based on his own account of the history of the double helix [9]. When Watson saw the X-ray diffraction pattern of DNA from Maurice Wilkins (King's college London), he got interested in finding the *structure* of the DNA — that would be the only way to understand genes. Watson worked with Francis Crick at the Cavendish. Earlier, Wilkins had showed the *DNA* X-ray patterns to a theoretician (Alex Stokes) who said that the pattern must have come from a helix. So Watson and Wilkins were sure it would be a helix. But they thought it would be a *triple helix* because of the estimated thickness and density known to Wilkins.

Around the same time (1951) Linus Pauling at Caltech (an all-time great chemist) established the α-helix structure of the protein molecule. Pauling often worked on macromolecule problems by playing with models which looked like preschool toys (made from balls, sticks, and glue). The success of this method inspired Watson to try a model building approach and hopefully prove that the DNA indeed was a helix. In the mean time Crick and Bill Cohran (also at Cavendish) developed a theory for the X-ray diffraction patterns from helical structures (the Crick-Cohran-Stokes theory of helical diffraction) and verified that the theory was consistent with Pauling's α-helix and its X-ray pattern.

Watson and Crick soon built the triple helix model for DNA. Wilkins and his colleague Rosalind Franklin from King's college London visited them and argued that the triple helix was inconsistent with the water content found in DNA (according to X-ray patterns obtained by Franklin). This halted all efforts for a while.

At this time Watson learned that Erwin Chargaff (Columbia University) had shown earlier that the concentration of the bases A and T were the same in DNA samples. So were those of C and G. Crick was slowly learning that A and T might stick by hydrogen bonding at their flat surface and so might C and G. There were papers by Gulland and Jordan showing that there was lots of hydrogen bonding even at low DNA concentrations. By combining this with Chargaff's observation Crick realized that the DNA molecule might have the bases paired up this way.

Pauling also got interested in finding the DNA structure, and he too came up with a *triple* helix model! Watson quickly found flaws in the chemistry of the structure: it would make the DNA neutral rather than weakly acidic (as it had earlier been shown to be). Watson shared this message with Wilkins and Franklin during a visit. Wilkins also showed Watson the most recent X-ray pictures of DNA taken by Franklin and her student Gosling. These were great pictures of the B-form DNA taken with some meticulous effort, and it immediately became obvious to Watson that the molecule ought to have a helical structure. (They were studying two forms of DNA, the crystalline (A form) and paracrystalline (B form).) He could even deduce later that it implied 3.4 nm periodicity (Fig. 1).

Watson and Crick then decided to build models for the DNA helix again. This time they tried the *double helix* model first, the joke being that *all biological objects came in pairs* [9]. From the 1951 work of Alexander Todd (Cambridge, England) they knew that the backbone of the DNA molecule was very regular (today known to be the sugar-phosphate backbone). Watson and Crick first tried a model where like-bases stuck together (A with A, T with T, and so on) by hygrogen bonding. This wrong path was chosen because they were using a wrong chemical configuration for the bases called the enol form. The American crystallographer Jerry Donohue at Cavendish convinced them to use the so-called keto form in the models. When attempting this, Watson made the most crucial discovery that the base A in one strand had to pair with T in the other. Similarly C and G would have to pair. Such pairs are held together by hydrogen bonding, and furthermore have similar shape. The resulting double helix was verified to be correct stereochemically, in addition to being consistent with X-ray diffraction patterns. It was also consistent with Chargaff's earlier observation that some bases have identical concentrations in DNA. The resulting model was readily accepted by Wilkins, Franklin, and Pauling. *"A structure as pretty as that just had to be right!"*

Watson and Crick had won the race. Their paper announcing the double helix appeared in the journal *Nature* on April 25, 1953 — a one-page paper reporting one of the greatest discoveries of science! In 1962 when Watson was 34, he shared the Nobel prize for Physiology or Medicine with Crick and Wilkins.

Box B2. Story of how Nature's greatest coding mystery was cracked

Perhaps the earliest proposal that genes did their work by generating proteins came in 1941 from Beadle and Tatum at Stanford. They worked with mold which grew on bread and argued that X-rays create changes (mutations) in some genes, affecting the generation of certains proteins (enzymes, to be specific). About ten years later, Linus Pauling and Harvey Itano at Caltech had evidence, based on their work on hemoglobin proteins, that each protein might have an associated gene. They showed sickle cells were caused by one single change in the amino acid chain (see Fig. 14). Then the famous physicist George Gamow proposed many possible mappings from DNA to protein, but nothing worked for a while.

The prediction that there ought to be an intermediate RNA molecule between DNA and protein was made first by Watson even before the double helix was invented. From this arose the central dogma of biology (Sec. 3) which is often credited to Crick who did much to popularize it. In a 1955 private communication to the RNA tie club members (a club founded by George Gamow [10]) Crick suggested that there ought to be an adaptor molecule for every amino acid, later found to be the tRNA. But the way in which it turns DNA into protein was not clear. In 1959 an enzyme called the RNA polymerase was discovered. It was involved in the production of single stranded RNA from double stranded DNA. The great moment came when the ribosome was discovered at the Massachusetts General Hospital, Boston. Here Paul Zamecnik was studying cell-free protein synthesis and could track amino acids radioactively. He found that they were being strung together at the sites of small molecules in the cell today known as the ribosomes. Zamecnik then worked with Mahlon Hoagland and showed that before these amino acids were assembled into a chain at the ribosome they were attached to some small RNA molecules. Watson and Crick pointed out that these ought to be the adaptors they were looking for, today known as the transfer RNA or tRNA molecules. The messenger RNAs (mRNAs) were verified to be the templates for proteins synthesis only in 1960. Details of the complete story (starting from the DNA through mRNA to protein) was worked out at Harvard, Caltech, and Cambridge (Watson, Matt Meselson, Francois Jacob, and Sydney Brenner).

The code that translates portions of DNA into specific sequences of amino acids came up next. Since there are only four choices for bases in DNA, a single base is not enough to specify one out of 20 amino acids. A sequence of three consecutive bases has $4^3 = 64$ combinations, so Sydney Brenner proposed that the trancription from the 4-letter DNA to the 20-letter protein takes place through triplets of bases (now called codons), each triplet specifying one amino acid. In 1961 Brenner and Crick at the Cambridge Labs then proved this experimentally, by deleting or inserting base pairs in DNA and seeing the effect on the resulting amino acid sequences. This was the first experimental proof of the existence of codons. The ability to force artificial mutations (insertion, deletion and alteration of bases) was crucial to these experiments. Also crucial was the fact that protein synthesis could be performed outside the cell in a test tube using a good supply of ribosomes, amino acids, transfer RNAs and a source of energy. Such a system would manufacture the proteins that correspond to an mRNA introduced into the test tube.

In 1961 Marshall Nirenberg from the National Institute of Health revealed at a conference in Moscow that the triplet *TTT* produces the amino acid phenylalanine (Phe or F). He found this by using the RNA molecule *UUUUUU....* (called poly-U) in a cell-free synthesis of amino acids. Thus 1/64th of the genetic code had been cracked. There still remained 63 triplets of bases for which the resulting amino acids had to be found out. This was completed in 1966 by Gobind Khorana at U. Wisconsin, and the complete genetic code had been cracked! The results were presented at the 1966 Symposium on gentic code in Cold Spring Harbor. The Nobel prize for this work went to Khorana, Nirenberg and Holley in 1968.